W130–W136 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 29 May 2014*
doi: 10.1093/nar/gku471

PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data

Daogang Guan[1], Jiaofang Shao[1], Zhongying Zhao[1], Panwen Wang[2], Jing Qin[2], Youping Deng[3], Kenneth R. Boheler[4,*], Junwen Wang[2,5,*] and Bin Yan[1,4,*]

[1]Department of Biology, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China, [2]Department of Biochemistry and HKU-SIRI, The University of Hong Kong, Hong Kong SAR, China, [3]Department of Internal Medicine and Biochemistry, Rush University Medical Center, Chicago, Illinois 60612, USA, [4]Stem Cell & Regenerative Medicine Consortium, LKS Faculty of Medicine and Department of Physiology, The University of Hong Kong, Hong Kong SAR, China and [5]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

Received February 5, 2014; Revised May 10, 2014; Accepted May 13, 2014

ABSTRACT

Interactions among transcriptional factors (TFs), cofactors and other proteins or enzymes can affect transcriptional regulatory capabilities of eukaryotic organisms. Post-translational modifications (PTMs) cooperate with TFs and epigenetic alterations to constitute a hierarchical complexity in transcriptional gene regulation. While clearly implicated in biological processes, our understanding of these complex regulatory mechanisms is still limited and incomplete. Various online software have been proposed for uncovering transcriptional and epigenetic regulatory networks, however, there is a lack of effective web-based software capable of constructing underlying interactive organizations between post-translational and transcriptional regulatory components. Here, we present an open web server, post-translational hierarchical gene regulatory network (PTHGRN) to unravel relationships among PTMs, TFs, epigenetic modifications and gene expression. PTHGRN utilizes a graphical Gaussian model with partial least squares regression-based methodology, and is able to integrate protein–protein interactions, ChIP-seq and gene expression data and to capture essential regulation features behind high-throughput data. The server provides an integrative platform for users to analyze ready-to-use public high-throughput Omics resources or upload their own data for systems biology study. Users can choose various parameters in the method, build network topologies of interests and dissect their associations with biological functions. Application of the software to stem cell and breast cancer demonstrates that it is an effective tool for understanding regulatory mechanisms in biological complex systems. PTHGRN web server is publically available at web site http://www.byanbioinfo.org/pthgrn.

INTRODUCTION

Gene regulation of eukaryotic organisms is a very complex process that is mainly carried out by tight interactions between transcription factors (TFs) and DNA sequences in specific ways (activation or inhibition). The ability of TFs to regulate target genes is modified by post-translational protein–protein interactions (PPIs) among TFs, cofactors and other proteins or enzymes upstream of transcriptional gene regulation. A variety of post-translational modifications (PTMs), including protein phosphorylation, acetylation and ubiquitination have been implicated in transcriptional gene regulation (1–3). Emerging evidences have indicated that PTMs of proteins are involved in many biological processes or human diseases through controlling DNA-binding ability to modulate downstream gene expression (4,5). In addition, epigenetic modifications, such as histone methylation, have been documented as crucial elements for regulation of genome function through changing chromatin architecture without altering DNA sequences. The combinatorial action of PTMs, TFs, as well as epigenetic alterations constitutes the hierarchical complexity in gene regu-

*To whom correspondence should be addressed: Tel: +852 3411 5834; Fax: +852 3411 5995; Email: yanbinai6017@gmail.com
Correspondence may also be addressed to Junwen Wang. Tel: +852 2831 5075; Fax: +852 2855 1254; Email: junwen@uw.edu
Correspondence may also be addressed to Kenneth R. Boheler. Tel: +852 2831 5405;Fax: +852 3017 5581; Email: bohelerk@hku.hk

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

lation. Despite its importance, we have incomplete knowledge of these processes, which limit our understanding of complex regulatory mechanisms.

The advent of high-throughput technologies allows biologists to examine molecule interactions at different levels, and generates multi-dimensional Omics datasets. There are several technologies used to detect PPIs, like co-immunoprecipitation with display technology (6), tandem affinity purification (7) and yeast two hybrid (8). Mass spectrometry-based methods can identify numerous PTMs (9), providing more insights into the associations between PTMs and gene expression. Coupling chromatin immunoprecipitation with next-generation sequencing (ChIP-seq) offers high resolution mapping of TFs or epigenetic modifications's interaction sites to genomic locations (10). ChIP-based studies detect the binding regions of these regulators on DNA sequences and show a genome-wide binding affinity between protein–DNA sequences. Therefore, ChIP-seq is informative for transcriptional or epigenetic regulatory relationships and reconstructing gene regulatory networks (GRNs).

A major goal of mining high-throughput data is to find underlying structures in the data that provide a basis for identification of regulatory modules and reconstruction of complex protein/gene networks. Computational algorithms have proven efficient for addressing this issue. For example, MINDy facilitates genome-wide identification of PTMs of TFs, and was successfully applied to determine the regulation of MYC activity in human B lymphocytes (11). Other methodologies have been developed to assemble both gene expression profiling and ChIP-seq or TF binding data for identifying TF–gene interaction modules, such as Bayesian multivariate modeling (12), matrix decomposition (13) and regression model (14). Graphical Gaussian model (GGM) was widely used for inferring GRNs by exploiting high dimensional throughput data (15–17). Recently, an integrative model Active Protein-Gene (APG) based on linear GGM with matrix decomposition was designed to connect upstream protein–TF networks with downstream TF–gene networks (18). Partial least squares (PLS) is a well-known regression tool suitable for statistical analysis of genomic and proteomic data, and modeling of gene networks and TF activity (19). Some studies have applied the PLS method to examine connections between TF or microRNAs and gene expression (20,21). Therefore, a strategy combining the GGM with PLS regression could be powerful way to develop network-related web software.

Online tools or servers have been reported that build transcriptional or epigenetic regulatory networks, for instance, ChIP-Array (22), ChEA (23) and RENATO (24). Our web-based framework CMGRN was devised for integrative analysis of causal relationships between regulators and complex gene regulation mediated by TFs, epigenetic factors and microRNAs (25). MAGNET server can generate and score PPI networks and coexpression gene–gene networks but did not connect both together computationally (26). Although these methods show improved results in uncovering transcriptional regulatory programs, there is a lack of effective web-based software capable of constructing underlying interactive networks linking post-translational and transcriptional regulatory components.

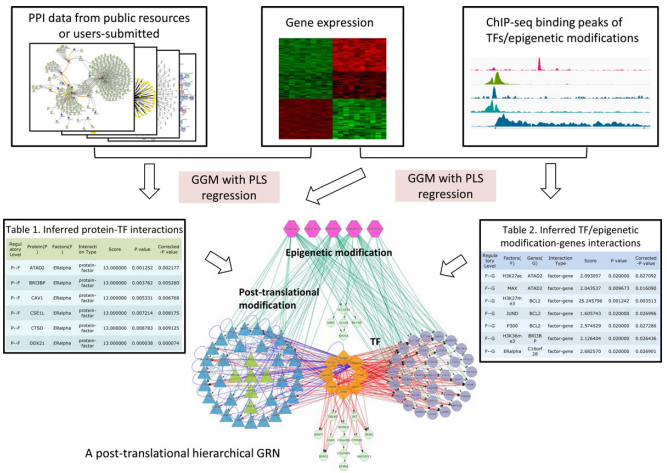

Figure 1. A workflow of integrated data analysis and GRN construction. PPI, TF binding and gene expression data are used to infer protein–TF interactions. The ChIP-seq binding peaks of TFs/epigenetic modifications are integrated with gene expression profile to infer TF/epigenetic modification–gene interactions. Finally, the tabulated results are organized to hierarchical GRNs.

A large number of PPI (such as BioGRID, STRING, HPRD and Reactome), ChIP-seq and gene expression (such as ENCODE, modENCODE, GEO and ArrayExpress) resources are publically available. To provide an easy bioinformatics tool to use and interpret the high throughput Omics data, we present an integrative web server, post-translational hierarchical gene regulatory network (PTHGRN) (http://www.byanbioinfo.org/pthgrn) to unravel relationships among PTMs, TFs, epigenetic modifications and gene expression. The newly developed server performs a GGM with PLS regression-based methodology to generate and score potential interactions of both protein–TF and TF/epigenetic modification-gene. It is system-wide and enables biologists to process the mixture information from PPI, ChIP-seq and gene expression, using standard data formatted with minimal need of bioinformatics skills. At the end, PTHGRN server robustly constructs hierarchical GRN for further evaluating the effect of PTMs on transcriptional regulatory complexity. Application of the software in mouse embryonic stem (ES) cell and human breast cancer demonstrates that it can explore biologically meaningful regulatory networks. Proof of principle analyses validates the use of PTHGRN for systems biology applications.

MATERIALS AND METHODS

Workflow of PTHGRN

As outlined in Figure 1, PTHGRN mainly conducts two tasks simultaneously. To examine how proteins (including cofactors, TFs, enzymes, etc.) affect TF performance post-translationally, input datasets, comprised of PPI, TF binding and gene expression, are submitted to the server. In this step, a GGM with PLS regression-based method will profile protein effects on TF activity and evaluate all interactions between TFs and proteins. The result is a list of the inferred protein–TF interactions that satisfy cutoff *P*-values (Table 1 of Figure 1), which can be used for constructing

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

upstream protein–TF networks. Another task is inference of regulator (TF/epigenetic modification)–gene interaction networks. Two types of input data are required: ChIP-seq binding peaks of the regulators or TF binding and gene expression. By performing the similar methodology, PTHGRN generates and scores all possible interactions between the regulators and genes. The resulting table is a list of the inferred regulator–gene interactions that satisfy cutoff *P*-values (Table 2 of Figure 1). Depending on availability of input data, users can carry out transcriptional regulatory network task with only gene expression and binding data of regulators. If with gene expression and PPI data only, the server will construct PPI networks, where the proteins are present in expression data. With all the three input data, the server organizes the identified protein–TF and regulator–gene interactions together and constructs a post-translational hierarchical GRN.

Scoring interaction networks

PTHGRN used a GGM (15–17) with PLS regression (19) to perform an integrated analysis of gene expression profiles with PPI and binding data of a set of regulators, and to discover protein–TF and regulator–gene interactions. The basic algorithm is to generate and then score every possible interaction. Based on the input data, the algorithm sets several matrices: E represents expression level of genes, T refers to concentrations of TF targeting genes, P stands for concentrations of proteins affecting TFs and TFA represents TFs activities under different conditions. Except the four matrices, the algorithm will also generate two temporary connectivity matrices (consisting of 1 and 0), A referring to one with TFs targeting genes and B referring to one with proteins modifying TFs. In the matrix A, '1' indicates that a TF target gene in the ChIP-seq or TF binding data is also found in matrix E, and '0' otherwise. Similarly, '1' in the matrix B indicates that a protein in input PPI is also present in the ChIP-seq or TF binding data, and '0' otherwise. To estimate initial observations of the matrix T, the median expression level of TF target genes from E and A are used. Similarly, initial observations of the matrix P is calculated from E and B based on the median strategy. According to the dependent and conditional independent features of Bayesian network, the algorithm can establish the joint distribution Pr(T, P, E, TFA) corresponding to the four matrices. A natural choice for representing continuous variables is the use of Gaussian distribution. According to the Gaussian model, if S is a node in graphical network, we have the conditional density of S given its parents:

$$\Pr(S|R_1,\ldots,R_n) \propto N\left(\sum_{ij}\delta_{ij}r_{t\in\{1,\ldots,n\}},\sigma^2\right),$$

where $R_1,\ldots,R_n$ are the parents of S; $N(\mu,\sigma^2)$ is the density function of the normal distribution with mean μ and standard variations σ, and $r_1,\ldots,r_n$ are the observations of $R_1,\ldots,R_n$ respectively. δ_{ij} is the effect strength of the ith variable on the jth variable. Since the activities of TFs depend on concentrations of both TFs and proteins, as described in APG (18), we can use δ_{ij} representing the interaction scores of TFs targeting genes or proteins modifying TFs to update the joint distribution Pr(T, P, E, TFA) above. The TFA is a hidden variable depending on concentrations of both TFs and proteins. To set the joint distribution parameters, we employed PLS regression (19) to find true TFAs and the associated interactions. PLS-based network component analysis offers computationally highly efficient and statistically robust strategies to identify likely true TFAs for any given connectivity matrix. In addition, it allows statistical assessment of the available connectivity information, and also the discovery of interactions and natural groupings among regulatory components. During the whole scoring process, we used the Maximum Likelihood Estimation method to obtain optimal parameters or network structure of the graphical model and maximize the probability of observed interaction scores (17,18).

Estimating probability of interaction networks

In order to select potential interaction networks, PTHGRN will evaluate P-values through a randomization test. First, the server will generate all possible interactions between protein–TF and regulator–gene based on the original data without data randomization, so called signal interactions. Next, to conduct the randomization test, the input data will be randomly permutated a given number of times. The number, or 'iteration', can be chosen in the input web interface, for example, iteration = 100, indicating that the server is going to run 100-time data permutations, i.e. 100-time tests. For every permutation, PTHGRN randomly picks up the same number of proteins as true protein–TF from the PPI database, and the same number of genes as true regulator–gene from the ChIP-seq or TF binding data. The server then randomly sets the parameters for the matrices with the same structure as signal interaction generation for every test. Following the same procedure of scoring interaction networks described above, the server will score each of the randomized interaction networks, so called background interactions, to form background scores. PTHGRN then compares the obtained score with the background scores, and calculates the P-values of every signal interaction. Furthermore, we corrected the P-value based on False Discovery Rate using R function P-adjust. The networks with P or corrected-P-value satisfying a certain cutoff (for example, $P < 0.05$ or corrected-$P < 0.05$) will be chosen for the lists in output tables and visualization. In general, the score of the likely true signal is higher than its corresponding backgrounds due to data randomization, indicating lower P-values. In the web interface, users can set cutoff of P-value or corrected-P-value for selecting the output interactions. The randomization-based method has been widely used in many computational biology studies (13,27–29).

Web interface

PTHGRN provides an easy web interface platform for incorporating three types of input data. The first one is a tab-delimited list of PPI derived from public databases BioGRID, STRING, Dip, HPRD, Intact, Mint and Reactome, and consists of two columns containing pairs of interacting proteins. Currently, PTHGRN contains PPI data

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

that covers human, mouse, rat, *Drosophila melanogaster* and *Caenorhabditis elegans*. The server system also supports uploading of user datasets with a similar format. For example, results from proteomics or metabolomics experiments could be treated as PPI input. The second one is tab-delimited regulatory signal, and consists of two columns including the name of regulators (TFs or epigenetic modifications) and their target genes. A regular method to prepare the signal data is to analyze significantly enriched ChIP-seq binding peaks of the regulators along genomic sequences by using MACS (30) or other ChIP-seq analysis software. ENCODE project stores available ChIP-seq binding peaks of the regulators. We mapped these binding loci to the genomic regions from transcriptional start site 2000 bp to transcriptional end site 1000 bp on human or mouse genome. We extracted these binding data, including 171 and 47 TFs, as well as 45 and 13 epigenetic modifications in human and mouse, respectively. Overall, our server recruited a total 135 cell types of ENCODE, 108 from human and 27 from mouse. We also collected ChIP-seq binding peaks of *D. melanogaster* (including 42 TFs and 28 epigenetic modifications) and *C. elegans* (including 91 TFs and 91 epigenetic modifications) from public database modENCODE. Currently, PTHGRN provides ready-to-use binding peaks of the regulators from the two projects. Moreover, users can upload their own binding data in the required format. For example, conserved binding motif of TFs on the promoter can be treated as regulatory signals. In the server, we provide TF target genes containing conserved binding sites on promoters of human, mouse and rat (retrieved from ECRbase, http://ecrbase.dcode.org/), as well as *D. melanogaster* (22). The last input is tab-delimited numeric matrix and gene expression data, where the columns are sample ID and each row represents gene names. The expression data should consider the same biological materials as ChIP-seq experiments if using ChIP-seq binding data. We suggest that expression data contain differentially expressed genes, and are separated to up-regulated and down-regulated subgroups, respectively. In all the three input files, the gene and protein ID format should be consistent for data integration. The current databases of PTHGRN support official gene/protein symbols.

The web interface provides example data from mouse ES cell line V6.5 and human breast cancer cell line MCF-7. The ES cell data contains mouse PPI, ChIP-seq binding peaks of five histone methylations (H3K4me3, H3K27me3, H3K79me2, H3K36me3 and H4K20me3), four TFs (Oct4, Sox2, Nanog and Tcf3) and PRC2 component Suz12 from GSE12241 and GSE11724 of NCBI GEO database. The gene expression data was extracted from GSE3231 of GEO and characterized with a time-course profiling of genes differentially expressed from embryonic status to embryoid bodies. In the breast cancer example study, we processed human PPI, ChIP-seq binding peaks of five histone modifications (H3K4me3, H3K27me3, H3K36me3, H3K9me3 and H3K27ac) from GSE31755 of ENCODE project, and six TFs (ERα, CEPBP, FOXM1, GATA3, JUND and MAX), as well as cofactors p300 and CTCF from GSE32465 of ENCODE and GSE25684 of GEO. The gene expression data is based on time-course profiling of genes differentially over-expressed in the estrogen-stimulated breast cancer cell from accession number E-TABM-742 of the EBI database.

After uploading the input files and selecting certain cut-off parameters, users can start the PTHGRN procedure by click 'Submit' button. The interface displays three cut-off choices, *P*-value, corrected-*P*-value and iteration. A low *P*-value indicates a high degree of confidence that a protein/TF can be modified by another one or that a gene can be targeted by a regulator. Following the submission action, the server promotes an output interface showing resultant networks exploring post-translational, transcriptional and epigenetic regulatory hierarchies in graphical and tabular formats. Based on the two main output tables (see Figure 1), the server automatically generates a third table with a list of interacting protein pairs that can modify the TFs in the second table. This process can be carried out by directly searching the PPI database in PTHGRN server. The graphical interface would show a view of post-translational GRN including the three output tables above. Users can choose the interaction links of TF–protein or regulator–gene under post-translational, transcriptional and epigenetic levels. For example, users can click a TF of interest, and see an overview of all nodes (its target genes and proteins that modify it) and edges linking to this TF. The graphical topology generated enables users to manipulate the pan/zoom choice, search and retrieve text information showing score and *P*-values of all interactive TF–protein or regulator–gene for a selected node and its associated subgraph. In addition, the server can retrieve NCBI protein or gene information for the selected node.

Implementation

PTHGRN platform is modularly designed to allow other combination of regulatory networks, metabolic networks, pathway networks and drug-target networks. The web server is written in PHP with supporting scripts written in Pascal. Network display was created using Cytoscape web. PTHGRN does not have excessive computing requirements. In its current deployment and under typical server loading conditions, PTHGRN can finish processing a typical input dataset with ∼5–15 min. Optionally, users provide an Email address to receive a link for finished results if large datasets or large numbers of iteration are requested. The results will be stored for a month by default, and storage times can be longer or shorter depending on user's requirement.

EXAMPLE STUDY

Mouse ES cell differentiation

Orchestrated interplays among three pluripotency TFs Oct4, Sox2 and Nanog with other TFs, cofactors and epigenetic modifications are critical for controlling ES cell self-renewal and differentiation (31,32). We applied PTHGRN to investigate the complex regulatory associations in mouse ES cell. Figure 2 displays an output hierarchical GRN coordinated by PTMs, TFs and histone methylations. This result implies that a multi-level regulatory complex is involved in down-regulating genes that are related to transcriptional regulation, DNA

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

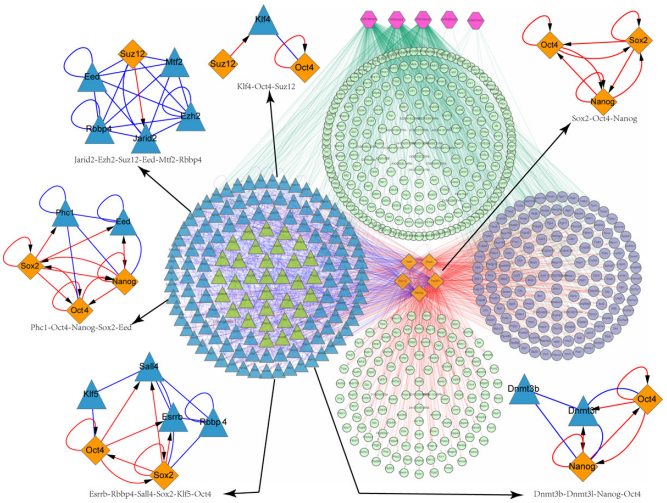

Figure 2. A post-translational hierarchical gene down-regulatory network in mouse ES cell. Triangle nodes represent proteins, with blue ones referring to those also targeted by TFs or epigenetic modifications and green ones referring to those only interacting with proteins. Blue non-directional edges are protein–protein interactions. Red and green arrow edges represent target genes of TFs and epigenetic modifications, respectively. Diamond and hexagon nodes represent TFs and epigenetic modifications, respectively. The six subgraghs represent the selected interactions among TFs, cofactors, epigenetic enzymes and other proteins in ES cell.

metabolic process and RNA processing, and thereby promoting ES cell differentiation. Under protein networks, we identified 44, 14 and 58 proteins that interact with the Oct4, Sox2 and Nanog respectively (Supplementary Table S1). In particular, several important modification interactions among TFs, epigenetic enzymes and cofactors were discovered (Figure 2), such as Sox2-Oct4-Nanog, Jarid2-Ezh2-Suz12-Eed-Mtf2-Rbbp4, Esrrb-Rbbp4-Sall4-Sox2-Klf5-Oct4, Phc1-Oct4-Nanog-Sox2-Eed, Klf4-Oct4-Suz12 and Dnmt3b-Dnmt3l-Nanog-Oct4, some of which are consistent with known reports (33–40). This finding validates that three core TFs Oct4, Sox2 and Nanog interact with each other and also with other TFs or proteins for the regulation of their downstream genes in stem cells (33). Three PRC2 components Ezh2 (H3K27 methyltransferase), Suz12 and Eed interact with each other to contribute to H3K27 methylation for repressing gene expression of mouse ES cells (37). Their cofactor Jarid2 can recruit the PRC2 proteins to common promoters for jointly regulating gene expression (36). Interestingly, ChIP-based experiments indicated that PRC2 binding is highly correlated with binding of Oct4, Sox2 and Nanog (41). Our analysis supports the roles of these key regulators during ES cell differentiation. At transcriptional level, Oct4, Sox2, Nanog can regulate 31–36% of genes differentially under-expressed respectively, where 102 are common targets. Their target genes include key TF or epigenetic enzyme genes, such as *Mybl2*, *Mycn*, *Sall4*, *Trp53*, *Esrrb*, *Eed*, *Ezh2*, *Phc1*, *Jmjd1a*, *Jmjd2c* and *Jarid2* (Supplementary Table S1). At epigenetic layer, we identified 41–43% of differentially under-expressed genes modified by three histone methylations, H3K4me4, H3K79me2 and H3K36me3 respectively. Epigenetic factors Suz12, H3K27me3 and H4K20me3 only regulate a few numbers of genes (Supplementary Table S1). Similarly, we tested differentially over-expressed genes using the same method. The result shows that repressive Suz12 and H3K27me3 are able to target 35 and 15% of up-regulated genes respectively, which are much greater than down-regulated ones (Supplementary Tables S1 and S2). This observation supports previous perspectives regarding role of PRC2 and the repressive epigenetic factors in leading to over-expression of developmental genes in the differentiated stem cells (42,43).

Estrogen-induced human breast cancer cell

Estrogen receptor alpha (ERα) is an estrogen-inducible TF that has been implicated in the development of human breast cancer. As a nuclear receptor, it can interact with many TFs, coregulators and growth factor-activated signalling to form a regulatory complex to modulate cancer-related biological processes (44). PTMs of ERα have been found through phosphorylation, ubiquitination, sumoylation and acetylation that affect its stability and activity (45). To dissect this complexity, we applied PTHGRN to analyze breast cancer data. The observation shows that ERα and cofactor p300 interacts with 73 and 33 proteins upon 24-h estrogen stimulation, respectively, where including TF proteins E2F1, FOS, MYC, FOXM1, IRS1 and IRS2 (Supplementary Table S3). ERα was found to link with other proteins to form protein networks, such as ERα-MYC-IRS2-IRS2-CAV1, ERα-MYC-HSPA8-HSPH1, ERα-MYC-TUBA1-TUBB, ERα-MYC-FOS-NCL and ERα-MYC-CTSD. The previous study has revealed that ERα and MYC physically interact with each other to stabilize the ERα-coactivator complex and to facilitate estrogen-mediated signaling networks (46). ERα also forms a complex with AP-1 family members (including FOS) to modulate bone-specific genes in osteoblasts (47). Although p300 and ERα does not directly interact each other, both can link with MYC and FOS (Supplementary Table S3). This analysis highlights a considerable interaction of ERα as a master player with many coregulators upstream of transcription in breast cancer cell. Furthermore, we identified target genes of eight TFs or cofactors and five epigenetic modifications (Supplementary Table S3). It is apparent that 25–27% of over-expressed genes could be modified by three active epigenetic factors, H3K4me3, H3K36me3 and H3K27ac, whereas repressive H3K9me3 and H3K27me3 only target a small number (4–6%) of genes. The epigenetic features of histone modifications could contribute to up-regulate estrogen-mediated gene expression in breast cancer. Taken together, the example in breast cancer confirms that ERα is a key factor of regulatory complex networks, which modulate estrogen-induced gene expression programs and promote molecular pathogenesis of breast cancer.

CONCLUSIONS

PTHGRN is a freely available web server for an integrated analysis of PPIs, ChIP-seq binding data and gene expression profiling. Using a GGM with PLS regression-based method, it can generate and score all possible interactions of protein–TF and TF/epigenetic modification-gene.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

Through statistical assessment, the server performs identification of highly potential interactions and reconstruction of hierarchical GRNs for evaluating roles of PTMs in transcriptional gene regulation. The web interface provides users ready-to-use input data derived from major high-throughput Omics resources, so that users have an option to submit these public or user-provided data. The workflow of PTHGRN produces output networks in graphical and tabular formats, which can be downloaded for further systems biology studies and visualization. The examples presented here using mouse ES and human breast cancer cells show the identified interaction networks to be biologically meaningful, in agreement with conclusions previously drawn from experiments and reports. It is expected that the newly developed server is able to provide a pilot framework for extensively unraveling regulatory associations among multilayered molecules in many biological complex systems, and thus would benefit not only biologists but also the bioinformatics community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Grant of Science Faculty of Hong Kong Baptist University [FRG1/13-14/008]; National Nature Science Foundation of China [91229105]; Hong Kong Research Grants Council [781511M]; Research Grants Council of Hong Kong Theme-based Research Scheme [T13-706/11]; Collaborative Research Fund of Hong Kong Research Grants Council [HKBU5/CRF/11G]. Funding for open access charge: Grant of Science Faculty of Hong Kong Baptist University [FRG1/13-14/008]; National Nature Science Foundation of China [91229105]; Hong Kong Research Grants Council [781511M]; Research Grants Council of Hong Kong Theme-based Research Scheme [T13-706/11]; Collaborative Research Fund of Hong Kong Research Grants Council [HKBU5/CRF/11G] .

Conflict of interest statement. None declared.

REFERENCES

1. Bae,S.C. and Lee,Y.H. (2006) Phosphorylation, acetylation and ubiquitination: the molecular basis of RUNX regulation. *Gene*, **366**, 58–66.
2. Brooks,C.L. and Gu,W. (2003) Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr. Opin. Cell Biol.*, **15**, 164–171.
3. Perkins,N.D. (2006) Post-translational modifications regulating the activity and function of the nuclear factor kappa B pathway. *Oncogene*, **25**, 6717–6730.
4. Gioeli,D. and Paschal,B.M. (2012) Post-translational modification of the androgen receptor. *Mol. Cell. Endocrinol.*, **352**, 70–78.
5. Xu,Y.M., Huang,D.Y., Chiu,J.F. and Lau,A.T. (2012) Post-translational modification of human heat shock factors and their functions: a recent update by proteomic approach. *J. Proteome Res.*, **11**, 2625–2634.
6. Li,M. (2000) Applications of display technology in protein analysis. *Nat. Biotechnol.*, **18**, 1251–1256.
7. Rohila,J.S., Chen,M., Chen,S., Chen,J., Cerny,R., Dardick,C., Canlas,P., Xu,X., Gribskov,M., Kanrar,S. *et al.* (2006) Protein-protein interactions of tandem affinity purification-tagged protein kinases in rice. *Plant J.*, **46**, 1–13.
8. Walhout,A.J. and Vidal,M. (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, **24**, 297–306.
9. Olsen,J.V. and Mann,M. (2013) Status of large-scale analysis of post-translational modifications by mass pectrometry. *Mol. Cell. Proteomics*, **12**, 3444–3452.
10. Furey,T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.
11. Wang,K., Saito,M., Bisikirska,B.C., Alvarez,M.J., Lim,W.K., Rajbhandari,P., Shen,Q., Nemenman,I., Basso,K., Margolin,A.A. *et al.* (2009) Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat. Biotechnol.*, **27**, 829–839.
12. Tang,B., Hsu,H.K., Hsu,P.Y., Bonneville,R., Chen,S.S., Huang,T.H. and Jin,V.X. (2012) Hierarchical modularity in ERalpha transcriptional network is associated with distinct functions and implicates clinical outcomes. *Sci. Rep.*, **2**, 875.
13. Yan,B., Li,H., Yang,X., Shao,J., Jang,M., Guan,D., Zou,S., Van Waes,C., Chen,Z. and Zhan,M. (2013) Unraveling regulatory programs for NF-kappaB, p53 and microRNAs in head and neck squamous cell carcinoma. *PLoS One*, **8**, e73656.
14. Geeven,G., van Kesteren,R., Smit,A. and de Gunst,M. (2012) Identification of context-specific gene regulatory networks with GEMULA–gene expression modeling using LAsso. *Bioinformatics*, **28**, 214–221.
15. Zhang,L. and Mallick,B.K. (2013) Inferring gene networks from discrete expression data. *Biostatistics*, **14**, 708–722.
16. Liu,Z.P., Zhang,W., Horimoto,K. and Chen,L. (2013) Gaussian graphical model for identifying significantly responsive regulatory networks from time course high-throughput data. *IET Syst. Biol.*, **7**, 143–152.
17. Ma,S., Gong,Q. and Bohnert,H.J. (2007) An Arabidopsis gene network based on the graphical Gaussian model. *Genome Res.*, **17**, 1614–1625.
18. Wang,J., Sun,Y., Zheng,S., Zhang,X.S., Zhou,H. and Chen,L. (2013) APG: an active protein-gene network model to quantify regulatory signals in complex biological systems. *Sci. Rep.*, **3**, 1097.
19. Boulesteix,A.L. and Strimmer,K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.*, **8**, 32–44.
20. Li,W., Zhang,S., Liu,C.C. and Zhou,X.J. (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.
21. Li,X., Gill,R., Cooper,N.G., Yoo,J.K. and Datta,S. (2011) Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. *BMC Med. Genomics.*, **4**, 44.
22. Qin,J., Li,M., Wang,P., Zhang,M. and Wang,J. (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res.*, **39**, W430–W436.
23. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma'ayan,A. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.
24. Bleda,M., Medina,I., Alonso,R., De Maria,A., Salavert,F. and Dopazo,J. (2012) Inferring the regulatory network behind a gene expression experiment. *Nucleic Acids Res.*, **40**, W168–W172.
25. Guan,D., Shao,J., Deng,Y., Wang,P., Zhao,Z., Liang,Y., Wang,J. and Yan,B. (2014) CMGRN: a web server for constructing multi-level gene regulatory networks using ChIP-seq and gene expression data. *Bioinformatics*, **30**, 1190–1192.
26. Linderman,G.C., Chance,M.R. and Bebek,G. (2012) MAGNET: MicroArray Gene expression and Network Evaluation Toolkit. *Nucleic Acids Res.*, **40**, W152–W156.
27. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
28. Guan,D.G., Liao,J.Y., Qu,Z.H., Zhang,Y. and Qu,L.H. (2011) mirExplorer: detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. *RNA Biol.*, **8**, 922–934.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

29. Li,M.J., Sham,P.C. and Wang,J. (2010) FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*, **26**, 2897–2899.
30. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
31. Ng,H.H. and Surani,M.A. (2011) The transcriptional and signalling networks of pluripotency. *Nat. Cell Biol.*, **13**, 490–496.
32. Keenen,B. and de la Serna,I.L. (2009) Chromatin remodeling in embryonic stem cells: regulating the balance between pluripotency and differentiation. *J. Cell. Physiol.*, **219**, 1–7.
33. Rizzino,A. (2009) Sox2 and Oct-3/4: a versatile pair of master regulators that orchestrate the self-renewal and pluripotency of embryonic stem cells. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **1**, 228–236.
34. Tanimura,N., Saito,M., Ebisuya,M., Nishida,E. and Ishikawa,F. (2013) Stemness-related factor Sall4 interacts with transcription factors Oct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. *J. Biol. Chem.*, **288**, 5027–5038.
35. Li,G., Margueron,R., Ku,M., Chambon,P., Bernstein,B.E. and Reinberg,D. (2010) Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev.*, **24**, 368–380.
36. Pasini,D., Cloos,P.A., Walfridsson,J., Olsson,L., Bukowski,J.P., Johansen,J.V., Bak,M., Tommerup,N., Rappsilber,J. and Helin,K. (2010) JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature*, **464**, 306–310.
37. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
38. Li,J.Y., Pu,M.T., Hirasawa,R., Li,B.Z., Huang,Y.N., Zeng,R., Jing,N.H., Chen,T., Li,E., Sasaki,H. *et al.* (2007) Synergistic function of DNA methyltransferases Dnmt3a and Dnmt3b in the methylation of Oct4 and Nanog. *Mol. Cell. Biol.*, **27**, 8748–8759.
39. van den Berg,D.L., Snoek,T., Mullin,N.P., Yates,A., Bezstarosti,K., Demmers,J., Chambers,I. and Poot,R.A. (2010) An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell*, **6**, 369–381.
40. Wei,Z., Yang,Y., Zhang,P., Andrianakos,R., Hasegawa,K., Lyu,J., Chen,X., Bai,G., Liu,C., Pera,M. *et al.* (2009) Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. *Stem Cells*, **27**, 2969–2978.
41. Lee,T.I., Jenner,R.G., Boyer,L.A., Guenther,M.G., Levine,S.S., Kumar,R.M., Chevalier,B., Johnstone,S.E., Cole,M.F., Isono,K. *et al.* (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, **125**, 301–313.
42. Bibikova,M., Laurent,L.C., Ren,B., Loring,J.F. and Fan,J.B. (2008) Unraveling epigenetic regulation in embryonic stem cells. *Cell Stem Cell*, **2**, 123–134.
43. Pasini,D., Bracken,A.P., Hansen,J.B., Capillo,M. and Helin,K. (2007) The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol. Cell. Biol.*, **27**, 3769–3779.
44. Mann,M., Krishnan,S. and Vadlamudi,R.K. (2012) Emerging significance of estrogen cancer coregulator signaling in breast cancer. *Minerva Ginecol.*, **64**, 75–88.
45. Anbalagan,M., Huderson,B., Murphy,L. and Rowan,B.G. (2012) Post-translational modifications of nuclear receptors and human disease. *Nucl. Recept. Signal*, **10**, e001.
46. Cheng,A.S., Jin,V.X., Fan,M., Smith,L.T., Liyanarachchi,S., Yan,P.S., Leu,Y.W., Chan,M.W., Plass,C., Nephew,K.P. *et al.* (2006) Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Mol. Cell*, **21**, 393–404.
47. Lambertini,E., Tavanti,E., Torreggiani,E., Penolazzi,L., Gambari,R. and Piva,R. (2008) ERalpha and AP-1 interact in vivo with a specific sequence of the F promoter of the human ERalpha gene in osteoblasts. *J. Cell. Physiol.*, **216**, 101–110.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015